Performance Metrics for the Evaluation of Hyperspectral Chemical Identification Systems

Eric Truslow[a], Steven Golowich[b], Dimitris Manolakis [b], and Vinay Ingle[a]

[a]**Northeastern University**
Department of Electrical and Computer Engineering
360 Huntington Avenue, Boston, MA 02115.
[b]**MIT Lincoln Laboratory,**
244 Wood Street,
Lexington, MA 02420-9185.

Abstract.

Remote sensing of chemical vapour plumes is a difficult but important task with many military and civilian applications. Hyperspectral sensors operating in the long wave infrared (LWIR) regime have well demonstrated detection capabilities. However, the identification of a plume's chemical constituents, based on a chemical library, is a multiple hypothesis testing problem which standard detection metrics do not fully describe. We propose using an additional performance metric for identification based on the so-called Dice index. Our approach partitions and weights a confusion matrix to develop both the standard detection metrics and identification metric. Using the proposed metrics, we demonstrate that the intuitive system design of a detector bank followed by an identifier is indeed justified when incorporating performance information beyond the standard detection metrics.

Keywords: Hyperspectral, LWIR, detection, identification, performance estimation.

1 Introduction

Passive hyperspectral sensors operating in the longwave infrared (LWIR) provide high resolution measurements in a region of the electromagnetic spectrum where many chemicals have unique absorption profiles. The high spectral and spatial resolution of these sensors allows for the identification of the individual chemicals within a gaseous plume.[1] A library of chemical absorption signatures is often all the prior knowledge available, and it is the job of the signal processing algorithms to decide if chemicals are present, and also which ones; the process of finding which chemicals are present in the plume is known as identification. A simple but effective system design consists of separate detection and identification algorithms, as shown in Fig. 1. In order to assess the performance of such a system, and to compare performance of different algorithms, it is necessary to define metrics that address both the detection and identification tasks. In this paper, we propose such a performance evaluation methodology based on confusion matrices. Furthermore, we employ this methodology to demonstrate that the design of a state-of-the-art detection algorithm followed by an identification algorithm is superior to that of either algorithm individually.

The difference between detection and identification is somewhat subtle. We categorize problems depending on the number of chemicals in the library and whether mixtures of chemicals can be present or not. These distinctions can be summarized as

Distribution A: Public Release. This work was sponsored by the Defense Threat Reduction Agency under Air Force contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the author and not necessarily endorsed by the United States Government.

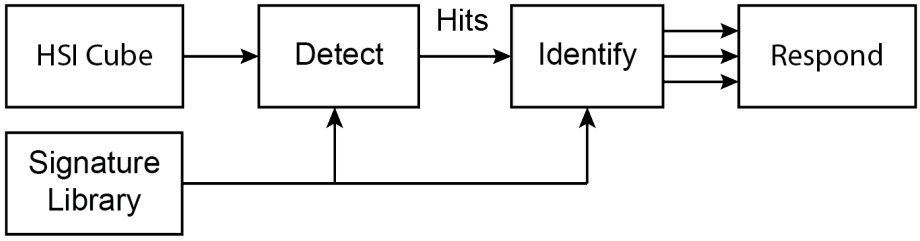


Fig 1: Practical chemical detection-identification system. The identifier has one output for each chemical in a known library.

1. Looking for a specific gas;
2. Looking for a single gas from a library of L gases;
3. Looking for mixtures of up to m gases from a library of L gases.

The first case is a detection problem where the library contains a single gas whose presence or absence needs to be decided; detection is inherently a binary hypothesis problem. When the library contains L chemicals and we are trying to determine which one is present, we no longer have a pure detection problem as there are multiple hypotheses to choose from. In Case 2, not only do we have the presence or absence of the plume to decide, but also which single chemical is actually present. In Case 3, instead of only picking a single chemical, a mixture of chemicals can be present.

Chemical identification can be formulated as a multiple hypothesis testing problem where each hypothesis represents a subset of the library gases, including the empty set. Each pixel in the scene has a true and an output hypothesis or class, which may differ. A natural way to represent the performance of such a system is through a confusion matrix, or error matrix, where each pair of truth and output is represented by a single entry of the matrix. Each entry of the matrix contains tallies of the number of pixels with the corresponding truth and output. A particular dataset and threshold produces a single realization of the confusion matrix, which can then be summarized in several useful performance metrics. For detection problems a single threshold determines the operating point and the confusion matrix can be used to estimate probability of detection (PD) and probability of false alarm (PFA). Sweeping a range of thresholds leads to a plot of PD versus PFA, called a receiver operating characteristic (ROC) curve. A ROC curve fully characterizes performance for detection problems.[2] In cases 2 and 3, the confusion matrix becomes larger and more difficult to interpret and in general may not be governed by a single threshold. When multiple thresholds are used, the construction of a ROC surface which characterizes performance is possible, but is not easily visualized and is still difficult to interpret.[3] The algorithms we consider use only a single threshold, considerably simplifying analysis. Evaluating the system for a range of thresholds then produces a performance curve for each metric used, from which general performance trends can be assessed.

The appropriate metrics to use in summarizing a confusion matrix depend on the particular application. In hyperspectral chemical detection and identification problems, the number of background (no gas) pixels is far larger than the number of plume pixels, making the false alarm rate one of the most important metrics of performance when operating a real system. Therefore, our approach partitions the confusion matrix into two parts: one involving the false alarm rate, and the

other involving the pixels that contain plume. For the portion of the confusion matrix containing plume pixels, we utilize two different metrics, the correct detection rate and the Dice index. The correct detection rate is the fraction of pixels for which at least one chemical is correctly detected. The identification metric we choose is known as the Dice index and is based on the amount of agreement between the truth and output hypotheses.[4,5] Unlike the correct detection rate, the Dice index considers both the number of correct chemicals in the output and the number of incorrect chemicals. Since identification deals with mixtures of chemicals and the correct detection rate does not incorporate mixtures, the Dice index is the metric we choose for evaluating identification performance. Ultimately, a good system should have: a low false alarm rate, a high correct detection rate, and high identification performance as measured by the Dice index.

The system we evaluate is a detector bank followed by an identifier. The detector bank is composed of adaptive coherence (cosine) estimator (ACE) detectors,[6] while the identifier is the Bayesian model averaging (BMA) approach.[7,8] Intuitively, using a bank of detectors as an identifier should perform worse as an identifier than an algorithm designed for identification; similarly, an identifier should perform worse at detection than a detector. We use the proposed identification metric and standard detection metrics to demonstrate that the ACE detector bank has better detection performance than BMA, but also has lower identification performance than BMA. These results suggest using the detector bank followed by the identifier for improved performance over either individually, which we demonstrate to be the case.

For single gas problems, algorithms can generally be categorized as classical detection algorithms or regression based algorithms. Classical detection algorithms include the matched filter, matched filter variants, the spectral angle measure (SAM), and the the adaptive cosine estimator (ACE).[2] Regression based algorithms fit the data to a regression model and then use either a significance test or a threshold on the model coefficients to determine whether or not a particular gas is present.[9] We choose the ACE algorithm as a detector since it is a very effective and popular detection algorithm for hyperspectral imagery.

Identifiers are designed for when mixtures of gases are permissible, and the constituents unknown. Several identification techniques have been proposed including using a bank of detectors,[6] using linear regression models with significance testing,[10,11] using step-wise regression,[12] and Bayesian techniques.[7,13,14] Compared to a detector, identifiers make a decision for each chemical in the library, whereas detectors only consider a single gas. BMA was selected for the identifier because it is considered state-of-the-art for identification.

To understand how performance is affected by the parameters of the plume, we used a plume embedding procedure to produce synthetic plume data that preserves the variability of the background data. Plume thickness has a non-linear relationship with the measured signal, which can be exploited by non-linear fitting algorithms, but causes additional fitting error in linear techniques like the ones we consider.[15] Both algorithms were tested individually on embedded data for a range of thickness parameters. The cascaded system was then tested for the same parameter ranges.

The key results of our study show that both the detector and the identifier are needed for overall good performance, a result that has not been demonstrated in the literature before. To our knowledge, using a confusion matrix to develop a series of performance metrics, and the use of the Dice index as a performance measure have not been done in this field. Our approach provides a starting place for comparative analysis of other system designs.

The remainder of this paper is organized as follows. Section 2 presents background material on the phenomenology of the data and explains the simplifications used in deriving useful models.

In Section 3, we discuss confusion matrices and the proposed identification performance metric. In Section 4, the two identification algorithms are defined, and the key formulas are presented. In Sections 5.1 and 5.2, we compare the detection and identification performance of the two identification algorithms individually. The effect of plume thickness on identification performance for each algorithm is explored in Section 5.3. Performance of the cascaded system with respect to plume thickness is examined in Section 5.4. Finally, in Section 6, we provide a short summary of the paper and discuss future work.

2 At-Sensor Radiance Signal Model

A simple but useful model for the at-sensor radiance in the longwave infrared (LWIR) can be developed from a full radiative transfer model with a few key assumptions:

1. The plume is optically thin and the distance between the plume and background are small enough to neglect the atmospheric transmittance in that region;
2. The plume is homogeneous in temperature and composition;
3. Scattering and reflections can be neglected.

These simplifications allow the use of the three layer model of Fig. 2 which can be used to derive our primary measurement equations using Kirchoff's law.[16] From Fig. 2 the measured radiance for an off-plume pixel is

$$L_{\text{off}}(\lambda) = (1 - \tau_a(\lambda))B(\lambda, T_a) + \tau_a(\lambda)L_{\text{b}}(\lambda)$$

where λ is in wavelengths or wavenumbers, T_a is the temperature of the atmosphere, τ_a is the transmittance of the atmosphere, L_{b} is the background radiance, and $B(\lambda, T)$ is the Planck function, which describes a black body at temperature T.

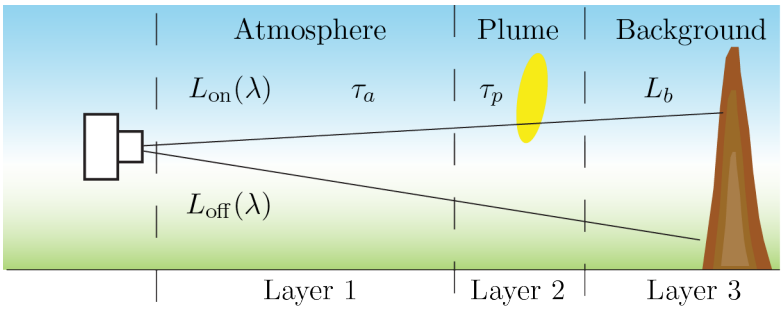


Fig 2: Simplified 3 layer radiance model for thin plumes.

The measured radiance for a pixel with plume becomes

$$L_{\text{on}}(\lambda) = (1 - \tau_a(\lambda))B(\lambda, T_a) + \tau_a(\lambda)(1 - \tau_p(\lambda))B(\lambda, T_p) + \tau_a(\lambda)\tau_p(\lambda)L_{\text{b}}(\lambda).$$

where τ_p is the transmittance of the plume. In terms of L_{off}, we instead have

$$L_{\text{on}}(\lambda) = (1 - \tau_p(\lambda))\tau_a(\lambda)(B(T_p, \lambda) - L_{\text{b}}(\lambda)) + L_{\text{off}}(\lambda). \quad (1)$$

The model in Eq. 1 is useful for analysis and gives a clear method for generating synthetic data under certain circumstances; namely, that the plume and atmosphere are in equilibrium.

The transmittance of the plume τ_p is governed by Beer's law:[16]

$$\tau_p(\lambda) = \exp\left(-\sum_{i=1}^{m} \alpha_i s_i(\lambda)\right) \tag{2}$$

where m is the number of gases in the plume, α_i is the concentration path length (CL) for gas i and s_i is the gas' absorption spectrum. When the thermal contrast $\Delta_T = T_p - T_b$ between the plume and background is small, and the background radiance is slowly varying with respect to wavelength, the difference $B(T_p, \lambda) - L_\text{b}(\lambda)$ is approximately proportional to Δ_T. Using the approximation $(1 - e^x) \approx x$, we obtain

$$x(\lambda) \approx \Delta_T \, \tau_a(\lambda) \sum_{i}^{m} \alpha_i s_i(\lambda) + L_\text{off}(\lambda) \tag{3}$$

which is linear in terms of the gas signatures. Assuming the atmospheric transmission $\tau_a(\lambda)$ is known, the signatures are multiplied by the atmospheric transmission τ_a before further processing is done. Ultimately, the temperature and CL both need to be estimated, but linear techniques estimate the product of the two. The problem of estimating the individual temperature and emissivity quantities is known as temperature-emissivity separation. For our purposes we estimate $\Delta_T \alpha_i$ as a single quantity b_i.

The input signal is convolved with the sensor response function and sampled at a set of band centers $[\lambda_1, \ldots, \lambda_p]$ to produce a measurement vector $\boldsymbol{x} = [x_1, \ldots, x_p]^\text{T}$ where p is the number of sensor channels. The atmospheric transmission $\tau_a(\lambda)$ is applied to library signatures $s_i(\lambda)$ and the product sampled to the sensor's resolution using the sensor's spectral response to obtain sampled signatures $\boldsymbol{s}_i$. Organizing the signatures as a matrix $\boldsymbol{S}$ and the b_i's as a vector $\boldsymbol{b}$, we have

$$\boldsymbol{x} = \sum_{i=1}^{m} \boldsymbol{s}_i b_i + \boldsymbol{v} = \boldsymbol{S}\boldsymbol{b} + \boldsymbol{v}, \qquad \boldsymbol{v} \sim \mathcal{N}\left(\boldsymbol{m}_b, \boldsymbol{C}_b\right) \tag{4}$$

where $\boldsymbol{m}_b$ is the background clutter mean and $\boldsymbol{C}_b$ is the clutter covariance. The assumption in Eq. 4 is that the clutter $\boldsymbol{v}$ is well modeled by a multivariate Gaussian distribution, which may not hold in reality, but is a useful model for many practical algorithms. Defining the whitening matrix $\boldsymbol{C}_b^{-1/2}$ and whitened vectors

$$\tilde{\boldsymbol{x}} = \boldsymbol{C}_b^{-1/2}(\boldsymbol{x} - \boldsymbol{m}_b), \qquad \tilde{\boldsymbol{S}} = \boldsymbol{C}_b^{-1/2}\boldsymbol{S}, \qquad \tilde{\boldsymbol{v}} = \boldsymbol{C}_b^{-1/2}(\boldsymbol{v} - \boldsymbol{m}_b) \tag{5}$$

yields the standard regression model

$$\tilde{\boldsymbol{x}} = \sum_{i=1}^{m} \tilde{\boldsymbol{s}}_i \tilde{b}_i + \tilde{\boldsymbol{v}} = \tilde{\boldsymbol{S}}\tilde{\boldsymbol{b}} + \tilde{\boldsymbol{v}}, \qquad \tilde{\boldsymbol{v}} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}\right). \tag{6}$$

where the clutter is zero mean and has identity covariance. The linear model of Eq. 6 is useful for developing the identification algorithms we consider, but is a good approximation only for thin plumes.

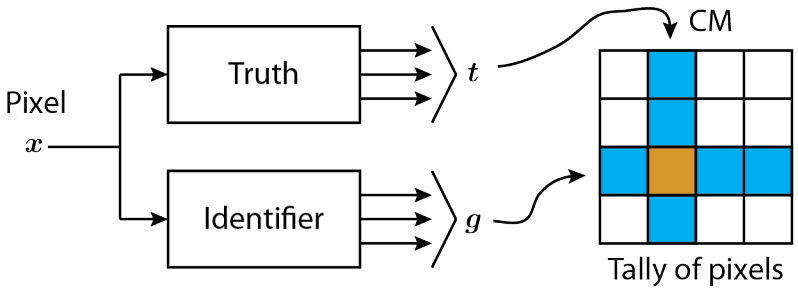


Fig 3: Algorithm independent way for confusion matrix construction.

3 Plume Identification Performance Metrics

The framework we present for evaluating identifiers relies on using the confusion matrix for multi-class problems. For a single dataset, the confusion matrix has all the performance information available in detail. However, for ease of interpretation, the confusion matrix may be summarized using several scalar performance metrics that are appropriate for detection and identification.

Assume that for each pixel we have the identifier output $\boldsymbol{g}$ and the "ground" truth $\boldsymbol{t}$. These are binary vectors given by $\boldsymbol{g} = \left[g_1, g_2, \ldots, g_L\right]^{\mathrm{T}}$ and $\boldsymbol{t} = \left[t_1, t_2, \ldots, t_L\right]^{\mathrm{T}}$ where g_k is an indicator for the kth gas being present in the identifier output, and t_k is an indicator for whether the kth gas is truly present in the pixel, i.e.

$$\mathrm{g}_k = \begin{cases} 1, & \text{Gas } k \text{ identified in pixel.} \\ 0, & \text{Gas } k \text{ not identified in pixel.} \end{cases} \qquad \mathrm{t}_k = \begin{cases} 1, & \text{Gas } k \text{ is present in pixel.} \\ 0, & \text{Gas } k \text{ is absent from pixel.} \end{cases}$$

The binary vectors $\boldsymbol{g}$ and $\boldsymbol{t}$ have M unique configurations depending on the maximum number of gases allowed in the output. The allowed configurations are denoted $\boldsymbol{g}_i$ and $\boldsymbol{t}_j$ with $i, j \in [1, \ldots M]$. Each possible truth vector $\boldsymbol{t}_j$ is assigned to a column of the confusion matrix (CM), while every possible identifier output is assigned to a row of the confusion matrix; each cell of the CM corresponds to a particular pair of truth and output vectors, and each pixel is assigned to a particular cell based on the vectors associated with it. In summary, the CM contains in element (i, j) a tally of the number of pixels with output $\boldsymbol{g}_i$ and truth $\boldsymbol{t}_j$. Operationally, the CM is constructed by tallying each pixel in the correct entry of the CM based on the identifier output and truth, as illustrated in Fig. 3.

The confusion matrix varies in size depending on both the size of the library and whether mixtures are allowed. In terms of hypothesis testing or classification, each hypothesis H_k has a corresponding binary indicator vector $\boldsymbol{g}$ or $\boldsymbol{t}$ depending on which library chemicals are present and on whether the hypothesis is the true one or the output from the system. When looking for only a single gas, the confusion matrix is only 2×2 as in Fig. 4a and a single threshold controls whether a pixel is assigned to the null-hypothesis (H_0) or the gas present hypothesis (H_1). When looking for one gas out of a library of size L, the CM is size $[L+1]\times[L+1]$ with the possibility of both correct identifications and incorrect identifications, as in Fig. 4b. Incorrect identification occur when one chemical is mistaken for another or when there is no overlap between the chemicals in the truth and output. In Fig. 4b, the hypotheses H_1 and H_2 represent each chemical from a library of size two; the corresponding binary vectors are $\boldsymbol{g}, \boldsymbol{t} = \left[1\ 0\right]$ or $\left[0\ 1\right]$. Hypothesis H_3 represents the presence

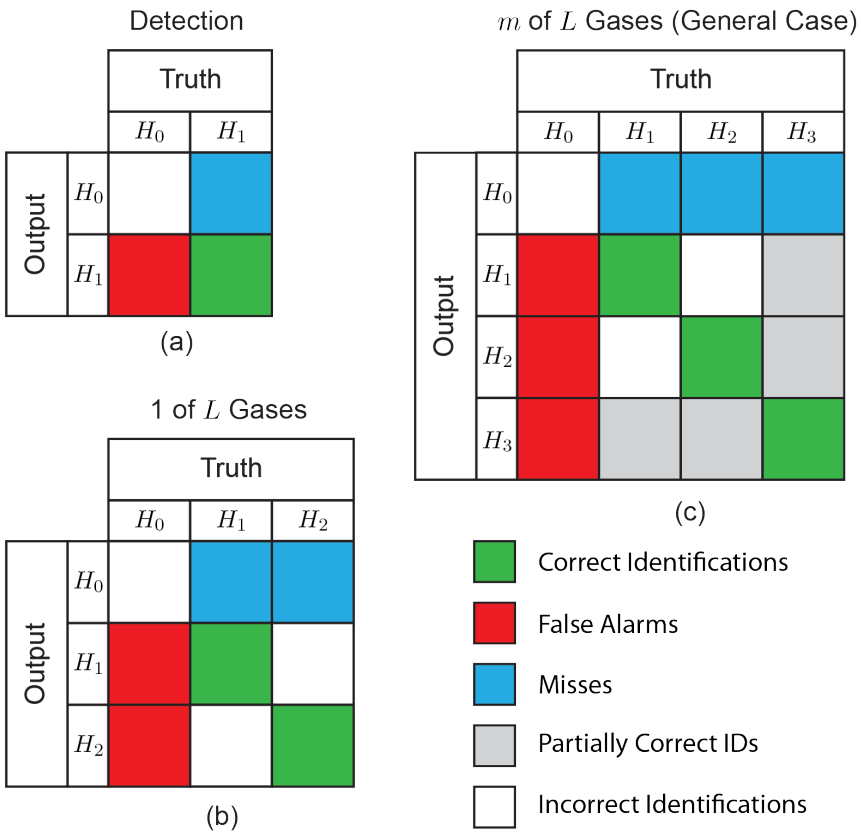


Fig 4: Different confusion matrices depending on the particular problem. (a) Detection type problem. (b) 1 of L gases with library of size 2. (c) Up to two gases with library of size 2. H_0 is the null-hypothesis; H_1 and H_2 contain gas 1 and 2 respectively, while H_3 has both.

of both chemicals in the plume. When looking for up to m of L the CM is of size $\sum_{k=0}^{m} \binom{L}{k}$, but in general when any mixture of chemicals is allowed the CM is size 2^L. In Fig. 4c, the full CM for a library of size two is shown; the hypothesis H_3 represents the mixture of both chemicals; when there is some overlap between the chemicals that are detected and the chemicals actually present we have a partially correct identification. In general, since the CM, and the number of hypotheses to test, grows exponentially with the size of the library, it is impractical to fill out the full confusion matrix or to test all possible models. For even moderately sized threat libraries the confusion matrix becomes difficult to interpret because of its size necessitating summarization.

The CM can be summarized by partitioning, weighting and then averaging. For plume identification applications the CM can be partitioned into several sub-matrices that contain: false alarms, misses, correct IDs, and incorrect IDs as shown in Fig. 4. Broadly, the false alarm section is the column where no gases are present ($\boldsymbol{t}_j = \mathbf{0}$) and the other cases occur when at least one gas is present. Performance metrics can be calculated using portions of the CM as follows: choose a partition the CM; sum the elements of the partition; apply a weight matrix $\boldsymbol{W}$ with weights $w_{i,j}$ to the CM; sum the weighted elements of the partition; take the ratio of the weighted and unweighted

β	$w_{i,j}$	Name(s)	Description
0	$\frac{(\boldsymbol{g}_i \cdot \boldsymbol{t}_j)}{\lvert\boldsymbol{t}_j\rvert}$	Sensitivity Recall	Fraction of gases in pixel that are detected.
1/2	$\frac{2(\boldsymbol{g}_i \cdot \boldsymbol{t}_j)}{\lvert\boldsymbol{g}_i\rvert + \lvert\boldsymbol{t}_j\rvert}$	Dice index	Incorporates both number of gases identified and number actually present.
		F-metric	Harmonic mean of precision and recall.
1	$\frac{(\boldsymbol{g}_i \cdot \boldsymbol{t}_j)}{\lvert\boldsymbol{g}_i\rvert}$	Precision	Fraction of gases detected that are present in the pixel.

Table 1: Weights derived from Eq. 8 using different values of β.

sums. These operations can be written succinctly as

$$d_{\text{perf}} = \frac{\sum_{(i,j)\in\mathcal{S}} [\text{CM} \odot \boldsymbol{W}]_{i,j}}{\sum_{(i,j)\in\mathcal{S}} [\text{CM}]_{i,j}}, \quad 0 \leq d_{\text{perf}} \leq 1, \tag{7}$$

where $\odot$ denotes element-wise multiplication, and the set S represents the sub-matrix to sum over. The brackets [.] with subscripts indicate a single element of the matrix within the brackets.

Though there are a substantial number of different metrics that can be derived from the CM depending on the weights used, the metric we use for identification performance comes from the family of indexes defined by

$$w_{i,j} = \frac{(\boldsymbol{g}_i \cdot \boldsymbol{t}_j)}{\beta|\boldsymbol{g}_i| + (1-\beta)|\boldsymbol{t}_j|} \tag{8}$$

with $\beta \in [0, 1]$. The numerator in Eq. 8 indicates the number of agreements between the output and truth, while the denominator has the sum of the number of chemicals in the truth and the number in the output. The resulting weights incorporate the truth and output vectors to varying degrees depending on the value of β used. Setting $\beta = 1/2$ in Eq. 8 we obtain the Dice index

$$w_{i,j} = \frac{2(\boldsymbol{g}_i \cdot \boldsymbol{t}_j)}{|\boldsymbol{g}_i| + |\boldsymbol{t}_j|} \tag{9}$$

or F-metric, not to be confused with the F-test from linear regression.[17] We choose the Dice index for identification because it incorporates both the number of gases in the identifier's output, and the number of gases in the pixel. Other choices of β lead to metrics that weigh the importance of $\boldsymbol{g}$ and $\boldsymbol{t}$ in different proportions. For example, when $\beta = 0$ the number of incorrect outputs is not taken into account. Several common weights derived from Eq. 8 that are used in the literature are presented in Table 1.[18]

The three metrics we use are the false alarm rate, the correct detection rate, and identification performance as measured using the Dice index of Eq. 9. These detection and identification metrics are listed in Table 2, along with the partitions of the CM used in the weighting and summarization process. The false alarm rate is very important in both detection and identification systems since it determines how much background data will have to undergo additional scrutiny. In standoff systems, the vast majority of data does not contain plume and is only background data; having a

low false alarm rate and having the other metrics as high as possible are desirable system characteristics.

The methodology we have taken for evaluating performance does not depend on any particular identification system architecture, except that the system must produce a list of identified gases. The internal workings of an algorithm are not taken into account when using this approach. However, the algorithms we analyze in this paper use a single threshold to make decisions. At each threshold we obtain a single realization of the CM from which each performance metric is calculated.

Statistic	Weights Used	Portion of CM Used
False Alarm Rate	$w_{i,j} = \begin{cases} 1, & \lvert \boldsymbol{g}_i \rvert > 0 \\ 0, & \lvert \boldsymbol{g}_i \rvert = 0 \end{cases}$	$j = 0$ (Gas absent)
Correct Detection Rate	$w_{i,j} = \begin{cases} 1, & \lvert \boldsymbol{g}_i \cdot \boldsymbol{t}_j \rvert > 0 \\ 0, & \lvert \boldsymbol{g}_i \cdot \boldsymbol{t}_j \rvert = 0 \end{cases}$	$j > 0$ (Gas present)
Identification Performance	$w_{i,j} = \frac{2(\boldsymbol{g}_i \cdot \boldsymbol{t}_j)}{\lvert \boldsymbol{g}_i \rvert + \lvert \boldsymbol{t}_j \rvert}$	$j > 0$ (Gas present)

Table 2: Detection and identification performance measures.

4 Plume Identification Techniques

The two algorithms we compare are a detector bank approach and a model averaging algorithm. The detector bank has a set of single-gas (single-chemical) detectors, one for each library signature. Each detector produces a score for a single gas, which is then thresholded to make a decision about each gas. Similarly, model averaging produces a score for each gas which is then thresholded to make a decision. Of the popular algorithms, the operational similarity of these two makes comparative analysis simpler, and easy to interpret. In this section we give overviews of the adaptive coherence estimator (ACE) detector bank and Bayesian model averaging (BMA) identification techniques, and provide the relevant formulas for each.

4.1 A Detector Bank for Identification

Detection algorithms are designed to solve binary hypothesis problems with a known target signal. Perhaps the most common and well known algorithm is the matched filter.[2] The normalized matched filter (NMF) is a simple modification where the matched filter is normalized by the measurement length. The matched filter and NMF can take both positive and negative values depending on the orientation of the input signal relative to the signature. In the LWIR, the relative direction of the input and signature may depend on the thermal contrast Δ_T from Eq. 3, which may be positive or negative. To create a sign-insensitive detector the NMF can be squared to obtain the adaptive coherence estimator (ACE). We define ACE for the kth library signature as

$$y_k = \frac{(\tilde{\boldsymbol{x}}^{\mathrm{T}} \tilde{\boldsymbol{s}}_k)^2}{||\tilde{\boldsymbol{x}}||^2 \ ||\tilde{\boldsymbol{s}}_k||^2} = \cos^2(\tilde{\theta}_k). \tag{10}$$

where $\tilde{\theta}_k$ is the angle between the whitened signature $\tilde{\boldsymbol{s}}_k$ and the whitened pixel $\tilde{\boldsymbol{x}}$. We use the term ACE for the squared detector, though various terms are used in the literature.[19]

To use ACE as an identifier, a bank of detectors can be constructed, where each detector is tuned to a particular library signature. For detection of a specific gas, only one ACE detector is needed; for detecting one of L gases, a bank of L detectors can be used and the maximum taken; for the m of L problem, instead of taking the maximum, the outputs can be thresholded to obtain a list of gases.

When only the maximum of the detector bank is considered, mixtures are excluded from consideration, which can be problematic when mixtures are present in the data. Picking the maximum may be appropriate in applications where only a single target is allowed in any pixel; for example, in the reflective regions of the electromagnetic spectrum, the ground resolution can be small enough that only a single target can be in any particular pixel.[20] For gaseous plumes, we use the thresholding approach instead of taking a maximum and then thresholding.

4.2 Bayesian Model Averaging

Bayesian model averaging (BMA) is a technique for estimating parameters using the construction of a set of models that are fitted to the data. In our case, each model is a linear model for $\tilde{\boldsymbol{x}}$ using a unique subset of gases from the library $\boldsymbol{S}$. Specifically, model j is in the form of Eq. 6 but with a particular library subset in $\boldsymbol{S}_j$ and the estimate of CL$\times\Delta_T$ in $\boldsymbol{b}_j$. Each model M_j is defined as

$$M_j: \quad \tilde{\boldsymbol{x}} = \tilde{\boldsymbol{S}}_j \tilde{\boldsymbol{b}}_j + \tilde{\boldsymbol{v}} \tag{11}$$

where the index j refers to the model, and is a separate index from other sections. Defining A_k as the event that gas k is present, BMA computes the probability of the event A_k as the average over all models

$$p_k = \Pr\{A_k|\tilde{\boldsymbol{x}}\} = \sum_j \Pr\{A_k|M_j, \tilde{\boldsymbol{x}}\} \Pr\{M_j|\tilde{\boldsymbol{x}}\} \tag{12}$$

where M_j is the jth model being considered. The probability $\Pr\{A_k|M_j, \tilde{\boldsymbol{x}}\}$ is an indicator of whether or not gas k is in the model. The model probabilities can be calculated using Bayes' rule

$$\Pr\{M_j|\tilde{\boldsymbol{x}}\} = \frac{\Pr\{\tilde{\boldsymbol{x}}|M_j\} \Pr\{M_j\}}{\sum_i \Pr\{\tilde{\boldsymbol{x}}|M_i\} \Pr\{M_i\}} \tag{13}$$

where $\Pr\{\tilde{\boldsymbol{x}}|M_j\}$ is the likelihood of the data given the model, and $\Pr\{M_i\}$ are the prior probabilities of the models. Typically, the likelihood of the data depends on model parameters that make it difficult to find expressions for the likelihood. Instead, the likelihood can be approximated using the model's Bayesian information criterion (BIC) as

$$\Pr\{\tilde{\boldsymbol{x}}|M_j\} \approx \exp\{-\text{BIC}_j/2\}.$$

For linear regression models, as in Eq. 11, the BIC is

$$\text{BIC}_j = n \ln(\text{RSS}_j/n) + d_j \ln(n)$$

where n is the number of spectral bands, and d_j is the number of gases in model j. The first term in the BIC depends on how well the model fits the data, and the second term is essentially a model complexity penalty. The residual sum of squares (RSS) is defined as

$$\text{RSS}_j = \tilde{\boldsymbol{x}}^{\text{T}}(\boldsymbol{I} - \boldsymbol{P}_{\tilde{\boldsymbol{S}}_j})\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{x}}^{\text{T}}\boldsymbol{P}^{\perp}_{\tilde{\boldsymbol{S}}_j}\tilde{\boldsymbol{x}}$$

where $\boldsymbol{P}_{\tilde{\boldsymbol{S}}_j} = \tilde{\boldsymbol{S}}_j(\tilde{\boldsymbol{S}}_j{}^{\text{T}}\tilde{\boldsymbol{S}}_j)^{-1}\tilde{\boldsymbol{S}}_j{}^{\text{T}}$. The additional penalty on the model complexity in the BIC leads to smaller models being more likely. Only considering models up to a certain size (mixtures of m) further favors smaller models.

Finally, assuming all models are equally likely, and that the models are exhaustive, the model probabilities must sum to one, and Eq. 13 becomes

$$\Pr\{M_j|\tilde{\boldsymbol{x}}\} \approx \frac{\exp\{-\text{BIC}_j/2\}}{\sum_i \exp\{-\text{BIC}_i/2\}}. \tag{14}$$

Eq. 14 and Eq. 12 together define the probabilities of each model and each gas occurring respectively. In BMA the probability for each gas to be present in Eq. 12 is then thresholded to obtain a list of gases present in each pixel.

5 Performance Evaluation of Identification Algorithms

To evaluate and compare different plume detection and identification systems, ground truth for the dataset is a necessity. In data with real plumes, the spatial extent of the plume, the constituent gases, the concentrations, and temperature of the plume are typically unknown. To obtain a performance estimate with known plume parameters, a synthetic plume embedding technique was used. The embedding technique is based on Eq. 1 and requires a background-only cube and a signature library.[15,21]

We selected a background cube and used the plume embedding algorithm to produce synthetic plumes with known ground truth at specified concentration pathlength (CL) values. We used a library of eight gas signatures based on the spectral library described in.[22] At least three of the gases in the library have strong spectral features in the same wavelength region; two of these gases were selected for embedding. The signatures were normalized to their maxima prior to embedding, thus the CL values reported can be used to infer the approximate thickness of the plume from Eq. 1 and Eq. 2. The data had 128 channels with centers ranging from 7.6 μm to 13.5 μm; most of the signatures had appreciable absorption peaks over this range of wavelengths. The plume was embedded over a ground portion of the image where the embedding model is most appropriate. Based on an estimate of the background temperature, the plume was simulated to be about 10K colder than the background; the temperature difference made the plume easily detected.

To compare the ACE and BMA algorithms fairly, the embedding region was excluded from background mean and covariance estimates, which are substituted into Eq. 5. The inclusion of the plume in these estimates can lead to substantial performance degradation. Both techniques processed the entire cube and produced scores for each pixel and each gas. BMA considered mixtures of up to three gases.

In the following sections we present results for several experiments using the same embedding region for several embedding scenarios. In Sections 5.1 and 5.2, we consider a single chemical at a single CL of 0.027. and examine the distribution of ACE and BMA scores with respect to

threshold. In Sections 5.3 and 5.4, we examine system performance over a range of CLs and embed both a single gas, and two gases in the same plume region. In Section 5.4, we propose a cascaded system and examine performance for both the single-gas and two-gas embeddings over the same CL ranges as in the other sections.

5.1 Detection Performance

Our analysis presents several histograms that were created as follows. For a background pixel, if any score exceeds the threshold, the result is a false alarm; therefore, if the maximum score among all the gas scores exceeds the threshold, it is a false alarm. For a plume pixel, there is one gas that is the correct gas embedded, the other gases are incorrect or wrong. If any other gas is above the threshold, then the identifier has made a mistake; if the maximum score among incorrect gases is above the threshold, then a mistake was made. The histograms we present in the following sections used only these maxima. The goal is to illustrate how important the threshold is to the performance of the system, and to highlight the difference between good detection performance and good identification performance.

For perfect detection, there should be perfect separation between background and plume scores. As shown in Fig. 5, ACE separates the background (red) and plume (green) quite well. Comparatively, BMA does not separate the background and plume as well. In Fig. 5a the red histogram is composed of the maximum scores for each background pixel since only the maximum score needs to exceed the threshold for the pixel to be a false alarm. The background pixel scores are distributed close to 0, with very few exceeding a value of 0.1. The green histogram contains the scores for the correct gas only, and consists only of the pixels within the embedded plume region; the plume pixel scores are distributed away from 0 with very little overlap between the background and plume scores. The small overlap of the green and red histograms leads to the ACE bank achieving a high correct detection rate for a large range of false alarm rates, as shown in Fig. 6 in red.

In Fig. 5b the background scores (red) of BMA are not tightly distributed near 0, and span the full range of thresholds. There is significantly more overlap between the background and plume scores (green) of BMA than for the detector bank. The overlap between the distributions means that for any threshold, the number of false alarms and missed detections will be higher for BMA than for the ACE bank, as shown in Fig. 6 in blue.

The curves in Fig. 6 were constructed by sweeping a range of thresholds to produce a series of confusion matrices from which the performance metrics are computed. At any particular threshold, a single realization of the confusion matrix is obtained. At each threshold PD and PFA were estimated using the correct detection rate and false alarm rate with the partitions and weights of Table 2. From Fig. 6 the probability of detection for BMA is lower than ACE for any PFA. Since the ROC curve of the detector bank is above the one for BMA, the detector bank is a better detector for this dataset.

5.2 Identification Performance

Detection performance was measured using the background scores for each gas and the plume scores for the correct gas only. The outputs for the incorrect gases were neglected when considering the plume pixels, but for identification, the scores for the incorrect gases determine whether we made a correct identification or not. Since a single threshold is applied to each output of the identifier, scores for chemicals that are not actually present may exceed the threshold. Multiple

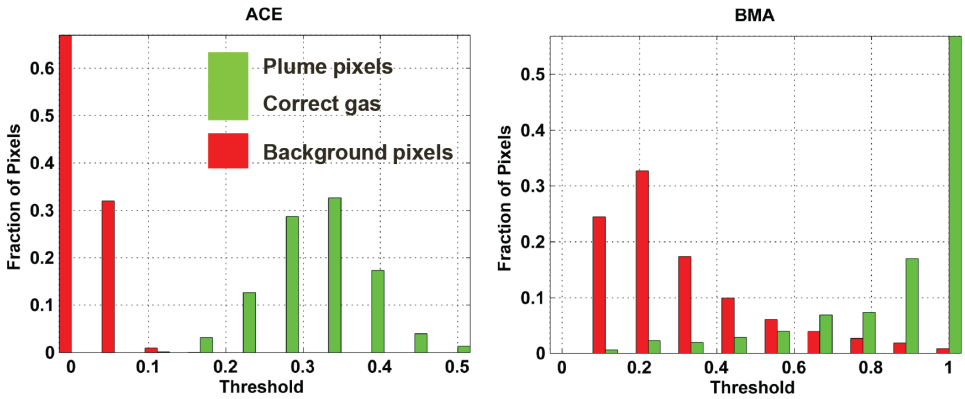


Fig 5: Scaled histograms of representative outputs for (a) ACE and (b) BMA using embedded data. Background and plume pixels are easily separated using ACE, but less so by BMA.

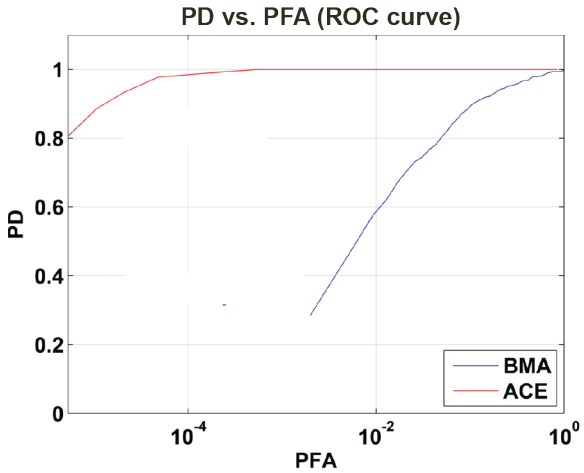


Fig 6: Receiver operating characteristic (ROC) curve for ACE and BMA based on the scores for the correct gas and for the background. The ROC curves can be constructed from the histograms in Fig. 5.

thresholds could be used, but selecting a threshold for each chemical individually is not practical without prior knowledge of which chemicals will be present. To have good identification performance, the distributions of scores for the correct library gas and the incorrect library gases should be well separated.

In Fig. 8, the same histograms as in the previous section are presented, but the scores for the incorrect gases are also included (blue). For each pixel in the plume, the maximum score among the incorrect gases was used to construct the histogram; since a single threshold is used to determine which gases were present, if the maximum exceeded the threshold, then an incorrect or partial identification occurred. If only the correct gas passes the threshold then a correct identification occurred. In Fig. 8a, the ACE detector bank shows good separation between the background and the plume for both the correct and incorrect gases; however, the correct gas and maximum incorrect gas histograms have significant overlap. In this case the ACE bank does a poor job of identifying exactly which gas is present when compared to BMA's results in Fig. 8b.

The identification performance curves shown in Fig. 9 were created by selecting a range of

thresholds, computing a confusion matrix for each threshold, and then using the partitioning and weighting scheme described in Section 3 with the Dice weighting of Eq. 9. The resulting performance curves demonstrate that ACE achieves a lower maximum than BMA; BMA has higher identification performance over a wide range of thresholds, achieving its maximum near a threshold of 0.5. The performance curves of Fig. 9 reflect approximately how well the blue and green histograms of Fig. 8 are separated. Overall, the performance curves indicate that BMA selects the correct gas more often than ACE for a wide range of thresholds.

We interpret these results as follows. With the chosen embedding parameters, several gases have similar ACE scores since each ACE detector considers only a single gas. The embedded chemical's signature is similar to two of the other library signatures, leading to multiple gases having similar ACE scores as indicated by the green and blue histograms of Fig. 8a; poor separation of different gases leads to poor identification performance when compared to BMA, as shown in Fig. 9. However, the ACE scores for the plume pixels are higher than for the background pixels, leading to good detection performance as indicated by the ROC curve in Fig. 6, and the green and red histograms of Fig. 5. In contrast, BMA achieves better separation of the correct and incorrect gases leading to higher identification performance for a wide range of thresholds.

The identification performance curve has an unusual characteristic compared to detection metrics, which monotonically increase or decrease with threshold. The identification performance curve is the average Dice score for the pixels within the plume at a number of thresholds. Performance initially increases with respect to threshold, reaches a maximum, and then decreases to zero at the maximum threshold. For small thresholds, several chemicals pass the threshold for a large portion of the plume pixels. In the low threshold regions, many of the plume pixels are partial identifications, while for high thresholds there are more misses as in the confusion matrices of Fig. 4c. The number of chemicals in the output of each pixel determines its Dice weight. For a library of 8 gases with 1 chemical embedded and 8 chemicals in the output, the weight is about 0.22, which is approximately the score BMA achieves at the lowest thresholds. For high thresholds, many plume pixels do not pass the threshold at all, leading to many pixels with a score of 0, causing the performance curve to deteriorate. Performance near one indicates that the majority of plume pixels have been correctly identified, i.e. all of the chemicals actually present are correctly identified and there are no extras.

Since the maximum identification performance of BMA is higher than ACE in this case, BMA can perform better as an identifier given an appropriate threshold. Selecting a threshold is usually accomplished by assuming that each threshold produces constant false alarm rate (CFAR). While it is possible in principle to select an operating point in this fashion, BMA has a much higher false alarm rate than ACE, as shown in Fig. 7.

Although BMA has a higher maximum performance than the ACE detector bank, the improved identification performance of BMA also comes with the undesirable higher false alarm rate. Chaining the two algorithms to have the detector followed by the identifier is an obvious way to have a lower false alarm rate while having the superior identification performance of BMA, as we discuss in Section 5.4.

5.3 Effects of Plume Thickness on Performance

The plume's thickness (CL) and temperature are the major drivers for how easily detected and identified the plume is. In the previous sections a single CL was used for embedding; in this section,

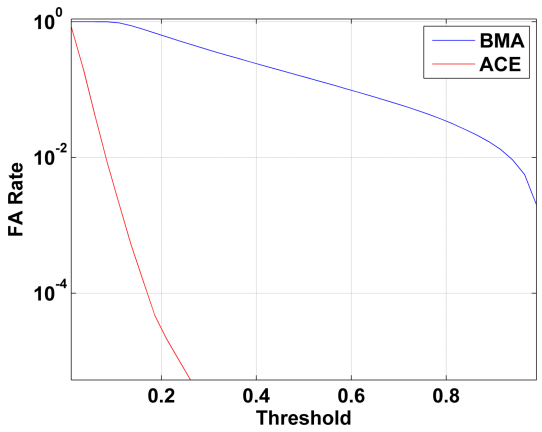
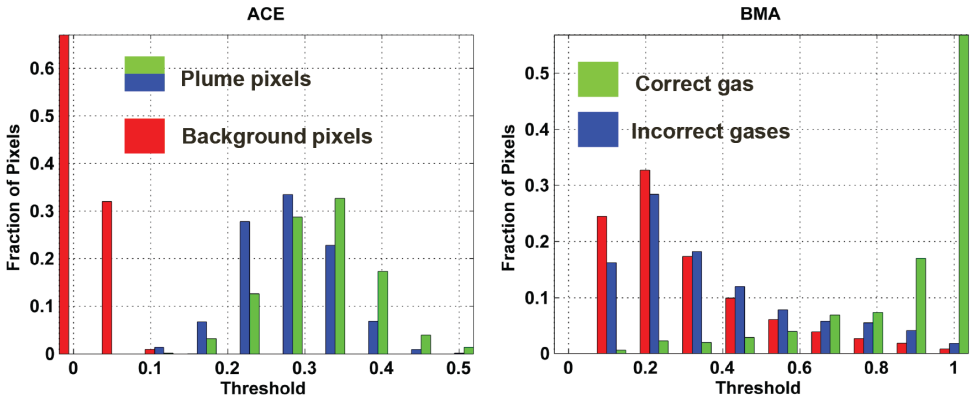


Fig 7: False alarm rates of the two identifiers vs. threshold.



(a) (b)

Fig 8: Scaled histogram of representative outputs for (a) ACE and (b) BMA using embedded data. Incorrect gases are not easily separated from the correct gas using ACE, and are better separated by BMA.

we examine identification performance for a range of CLs. We focus on the effects of CL instead of temperature because the measured signal is approximately linear in terms of the temperature contrast, as in Eq. 3. However, the non-linear relation between plume thickness and the measured signal is one of the reasons chemical identification is challenging. The same background cube and embedding region as in Section 5.1 were used for this section. First, a single gas was embedded for a range of CLs and both algorithms run on the data. For consistency, the same gas as Section 5.2 was used. Second, a mixture of two gases was embedded and the experiment repeated. The second gas used in the mixture was one that was spectrally similar to the first one. In both cases, the background statistics were the same for each CL. Consequently, at any particular threshold, the false alarm rate for each algorithm remained the same for all CLs.

Since we expected identification performance to deteriorate for sufficiently thick plumes, embedding was done for two different ranges of CLs. The first range simulated very thin plumes, while the second range was chosen to show identification performance reduction with thick plumes. Fig. 10a and Fig. 10c have maximum CLs of 0.1, which corresponds to $\max(\alpha s(\lambda)) = 0.1$ in Beer's law of Eq. 2 and the measurement equation Eq. 1. This corresponds to a minimum transmission of $\min(\tau_p) = e^{-0.1}$ Therefore, the minimum transmission for the plume is about 90%

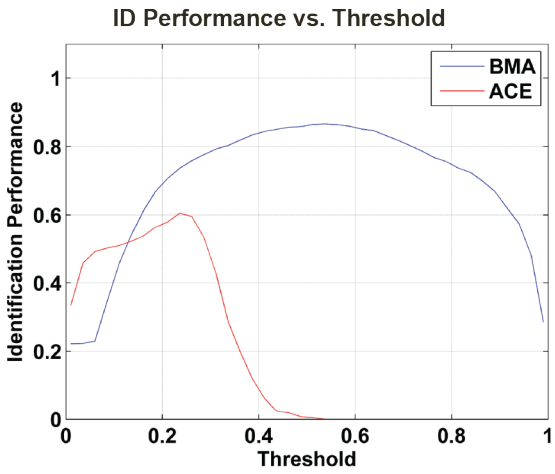


Fig 9: Identification performance for ACE and BMA over a range of thresholds.

in this region. The change in signal is approximately linear with respect CL in for the small CL region; however, in Fig. 10b and Fig. 10d, a larger range of CLs is shown. At a scaled CL of 10, the plume is almost completely opaque ($\tau_p \approx 0$) at its maximum absorption channel.

For the embedded plume with a single gas, the identification performance for both BMA and ACE for a range of CLs is shown in Fig. 10a. Performance is plotted with respect to CL for thresholds from 0.1 to 0.99 for BMA and from 0.1 to 0.9 for ACE. For most of the selected thresholds, BMA's performance increases with CL and is generally higher than ACE's performance for low CLs. After peak performance, BMA's performance slowly decreases, while ACE reaches a peak quickly and then decreases quickly. As the plume gets even thicker, as in Fig. 10b, the performance of BMA drops off from the maximum but ACE's performance has several peaks, depending on the threshold. The plume becomes too thick for BMA to distinguish individual gases and performance degrades substantially. However, with a very high ACE threshold, decent performance for very thick plumes can be achieved. Using ACE, it is possible to set a high enough threshold to separate the correct gas from the others when the plume is very thick. However, this sacrifices performance at small CLs, which is generally where standoff systems are expected to operate. For both algorithms, the general trend is that as CL increases identification performance increases, then reaches a peak, and then begins to degrade before completely failing. The poor performance of both algorithms for very thick plumes indicates that multiple gases have similar scores and cannot be separated by a single threshold, or that incorrect gases are being identified. When the plume becomes very thick, both techniques fail in identifying the plume and should not be used.

To test performance for mixtures, two gases were embedded in the same location as the previous experiment. The CLs of both gases were varied. In Fig. 10c, the performance of both algorithms is shown for a range of CLs. The trends are similar to the previous experiment except that performance of the detector bank is uniformly worse than BMA. Again, the problem ACE has in this case is that multiple gases have similar scores, and that the correct mixture of two gases is not considered by ACE. BMA gives both correct gases high scores because the model containing the mixture has a relatively high probability. The result is that BMA performs well for this mixture relative to ACE. In Fig. 10d, a similar degradation in identification performance is seen as with the single gas embedding. However, as compared to the single gas embedding, BMA performs better

for the mixture over a wider range of CLs than ACE.

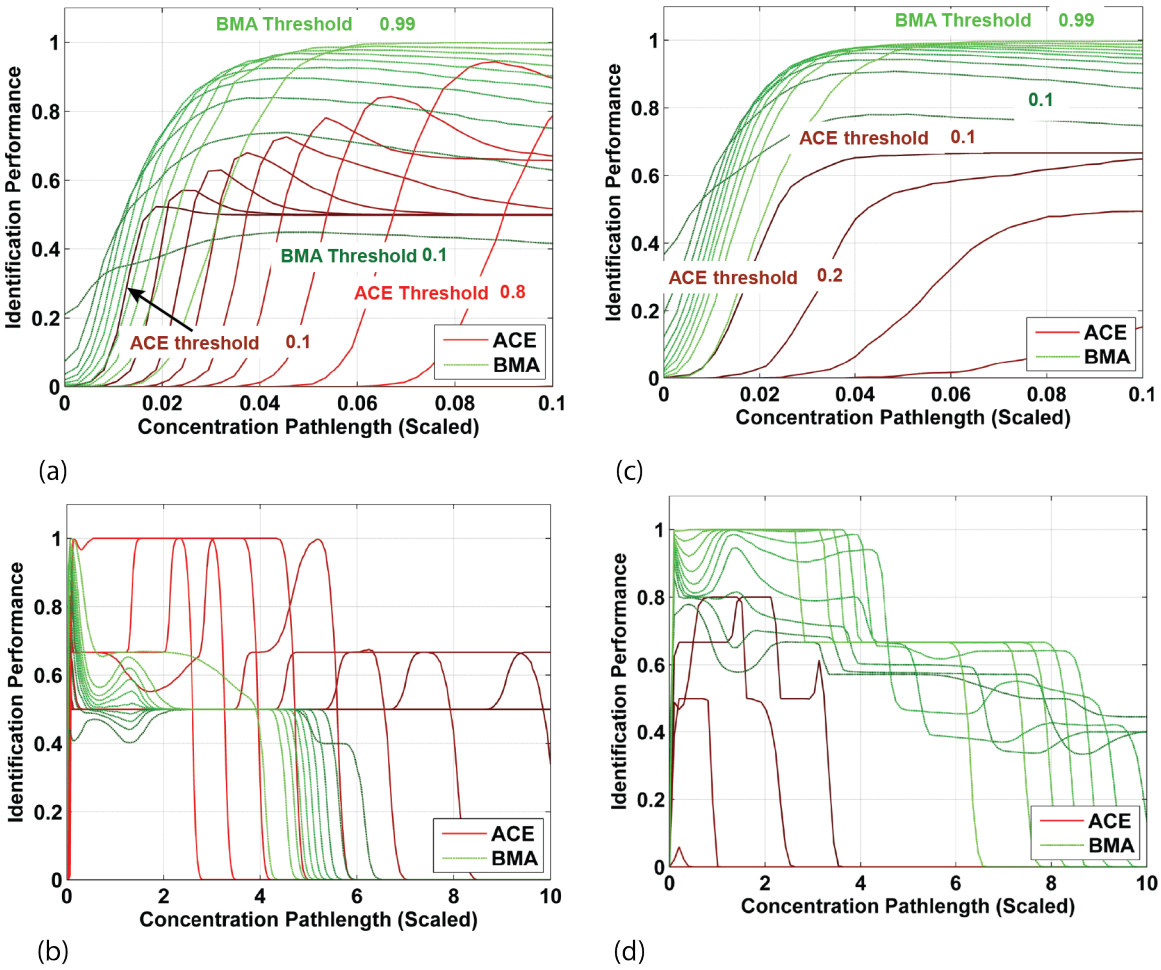


Fig 10: Identification performance of ACE and BMA with an plume at various concentration pathlengths (CLs) and thresholds. (a-b) Single gas embedded in cube. (c-d) A mixture of two gases embedded in cube. Thresholds were uniformly spaced between 0.1 and 0.9 for ACE, and 0.1 and 0.99 for BMA.

5.4 Detection Followed by Identification

The detection performance of the ACE filter bank is relatively high compared to BMA, but the identification performance of ACE is lower than BMA. We argue that combining the two algorithms in cascade leads to a system with superior performance characteristics compared to either algorithm individually. The cascaded system uses the ACE detector bank as a first pass and then passes only the hits to the BMA identifier and is shown in Fig. 11. Using the ACE bank as a first pass for the data can yield a high plume detection rate at a low false alarm rate. Each pixel that passes the threshold is then passed to BMA for identification, which makes a final identification decision about those pixels. In this section, the cascaded system is evaluated using the same embedding scenarios as before.

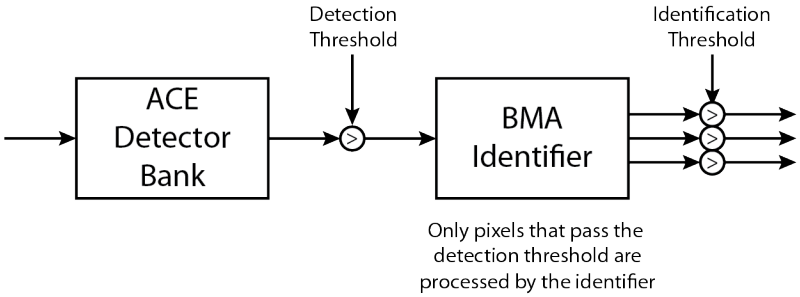


Fig 11: The cascaded system design.

The cascaded system was run on the embedded data using pairs of ACE and BMA thresholds. The number of false alarms for a particular threshold is constant with respect to CL, and is determined primarily by the ACE threshold. We selected two ACE thresholds: 0.1 and 0.36. The corresponding PFAs for both ACE and the cascaded system are 3×10^{-3} and zero; the second threshold is high enough so that no background pixels pass the ACE threshold. The system was tested at 10 evenly spaced BMA thresholds ranging from 0.1 to 0.99. The PFAs for BMA alone at these thresholds are 0.002 at the highest threshold and 0.99 at the lowest threshold.

The resulting identification performance curves for the single-gas embedding are shown for the low ACE threshold in Fig. 12a and for the higher threshold in Fig. 12b. The results when two gases were embedded are shown in Fig. 12c and Fig. 12d for the low and high ACE thresholds. The blue dashed curve shows the performance of the ACE detector bank alone; the solid green curves show BMA's performance alone; the red dotted curves are the cascaded system's performance.

The cost of cascading the two algorithms compared to ACE alone is generally worse identification performance at low CLs, which is more pronounced in Fig. 12d. However, at higher CLs, the cascaded system achieves better performance than the ACE bank for most choices of BMA threshold. Selecting the lowest BMA threshold of 0.1 actually results in worse performance than the ACE system alone for the single-gas embedding. Selecting the lower ACE threshold leads to a smaller difference between the green and red curves at smaller CLs but leads to a higher false alarm rate. In practice, the maximum operational false alarm rate will dictate what ACE threshold to select. However, it is unclear how best to set a BMA threshold in the cascaded system. From our experiments, the trends show that low thresholds lead to higher identification performance when the plume is very thin, but become comparatively worse as the plume thickens.

The results in Fig. 12d are a case where the combined system does substantially better than ACE alone because the highest ACE scores occur for a gas that is not present in the plume. In this case, the plume passes the threshold when the plume is sufficiently thick, but the correct gases are not detected or identified until the identifier. Instead, the mixture is incorrectly identified by the ACE detector as a completely different gas in the library. The detector is still performing the vital role of reducing the overall false alarm rate since using BMA alone has a substantially higher false alarm rate, as illustrated in Fig 7. The second pass by BMA correctly identifies the mixture more often and substantially improves identification performance. However, using ACE at the higher 0.36 threshold leads to degraded identification performance for thinner plumes.

Since the probability of false alarm for the identifier is substantially higher than the detector bank, the overall PFA should be set using the detector bank. It is tempting to set a threshold for the

identifier that maximizes identification performance; however, the maximum depends on the properties of the plume, which are known for synthetic data, but are unknown in real data. Even with good plume models, it is impractical to try and find a good threshold over all possible simulated scenarios. Based on our results, BMA thresholds greater than 0.5 showed decent performance over a range of CLs. The identification performance of the system is somewhat insensitive to the BMA threshold except at the extremes close to 0 and close to 1. How to select the identification threshold is an aspect for future work.

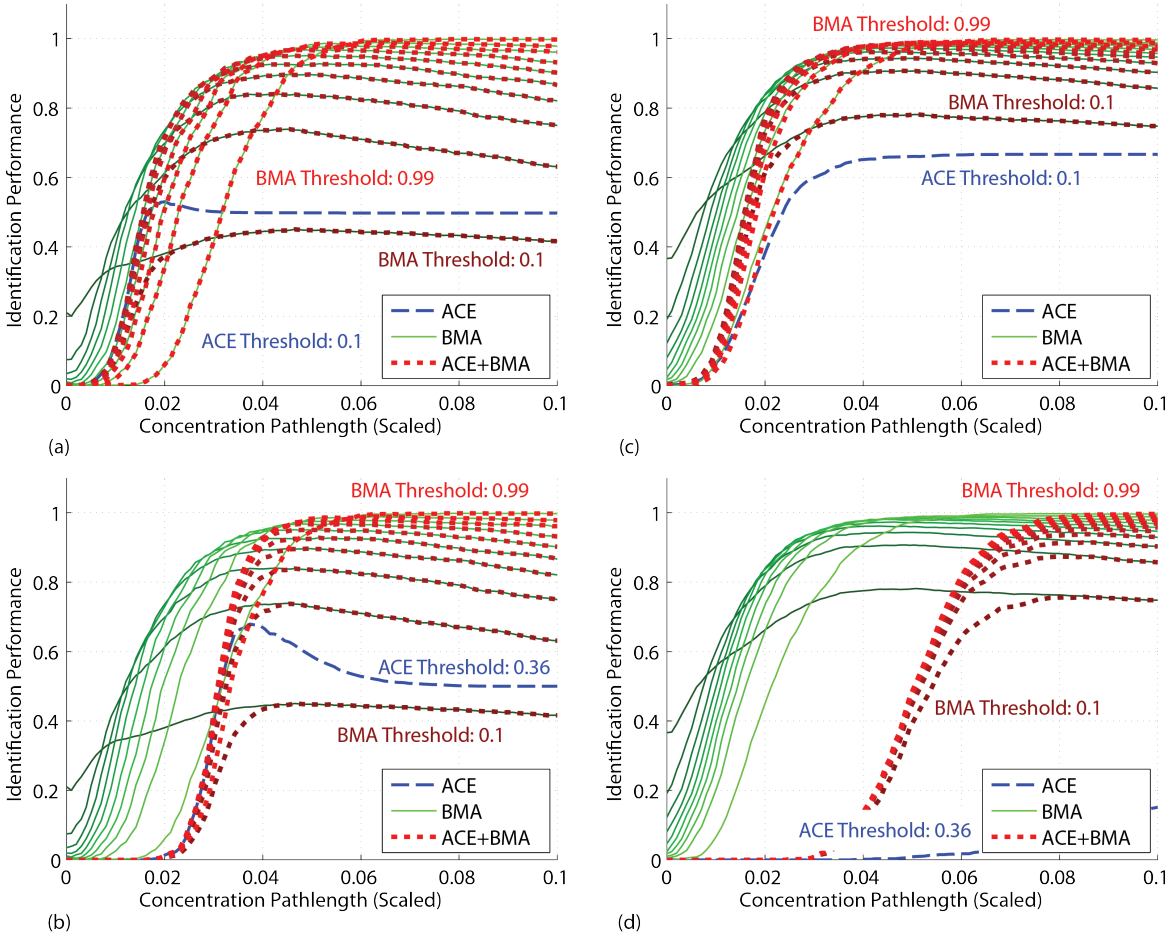


Fig 12: Performance of the cascaded system for BMA thresholds from 0.1 to 0.99. (a-b) Performance for the single-gas embedded data with ACE thresholds of 0.1 and 0.36 respectively. (c-d) Performance using the two-gas embedded data.

6 Conclusions and Future Work

The two main contributions of this work are the development of a performance metric for the evaluation of chemical plume detection and identification algorithms, and a demonstration that a detector followed by an identifier yields superior performance compared to using either alone. The approach to performance evaluation using a weighted confusion matrix and performance evaluation using the Dice metric are novel in this area of remote sensing. We applied our metric to quantitatively demonstrate that a cascaded detector and identifier can, attain a high PD and low

PFA, while also achieving a high ID score. Each constituent algorithm, by contrast, can achieve only one of these goals.

In the future, other types of algorithms should be evaluated, and the number of datasets expanded. Our study here is not exhaustive, but provides a framework place for further investigations. In this work, a single dataset with a single synthetic plume was used, but a range of concentrations and chemical mixtures should be incorporated into future work. As discussed previously, a study of all possible combinations of parameters is impractical; instead, smart experimental design can indicate what the overall trends are. In particular, a wider range of parameters and background clutter data will give insight into the problem of threshold selection for a cascaded system.

References

1 Manolakis, D., Golowich, S., and DiPietro, R., "Long-wave infrared hyperspectral remote sensing of chemical clouds: A focus on signal processing approaches," *IEEE Signal Processing Magazine* **31**(4), 120–141 (2014).

2 Manolakis, D., Truslow, E., Pieper, M., Cooley, T., and Brueggeman, M., "Detection algorithms in hyperspectral imaging systems: An overview of practical algorithms," *IEEE Signal Processing Magazine* **31**(1), 24–33 (2014).

3 Edwards, D., Metz, C., and Kupinski, M., "Ideal observers and optimal roc hypersurfaces in n-class classification," *Medical Imaging, IEEE Transactions on* **23**, 891–895 (July 2004).

4 Shen, X., Boutell, M., Luo, J., and Brown, C., "Multilabel machine learning and its application to semantic scene classification," *Proc. SPIE* **5307**, 188–199 (2003).

5 Dice, L. R., "Measures of the amount of ecologic association between species," *Ecology* **26**(3), pp. 297–302 (1945).

6 Tremblay, P., Savary, S., Rolland, M., Villemaire, A., Chamberland, M., Farley, V., Brault, L., Giroux, J., Allard, J., Dupuis, E., and Padia, T., "Standoff gas identification and quantification from turbulent stack plumes with an imaging Fourier-transform spectrometer," in [*Advanced Environmental, Chemical, and Biological Sensing Technologies VII*], *Proc. SPIE* **7673**, 76730H–12 (Apr. 2010).

7 Burr, T., Fry, H., McVey, B., and Sander, E., "Chemical identification using Bayesian model selection," in [*Proceedings of 2002 Spring Research Conference on Statistics in Industry and Technology*], (2002).

8 Burr, T., Fry, H., McVey, B., Sander, E., Cavanaugh, J., and Neath, A., "Performance of variable selection methods in regression using variations of the Bayesian information criterion," *Communications in Statistics - Simulation and Computation* **37**, 507–520 (Feb. 2008).

9 Young, S. J., "Remote sensing of CO2 in Industrial-Stack plumes using thermal infrared hyperspectral imaging," Tech. Rep. ATR-2003 (8407)-3, Aerospace Corporation (2003).

10 Harig, R., Matz, G., and Rusch, P., "Scanning infrared remote sensing system for identification, visualization, and quantification of airborne pollutants," *Proc. SPIE* **4574**, 8394 (2002).

11 Harig, R. and Matz, G., "Toxic cloud imaging by infrared spectrometry: A scanning FTIR system for identification and visualization," *Field Analytical Chemistry & Technology* **5**(1-2), 75–90 (2001).

12 Pogorzala, D., Messinger, D., Salvaggio, C., and Schott, J., "Gas plume species identification in airborne LWIR imagery using constrained stepwise regression analyses," in [*Defense and Security*], *Proc. SPIE*, 194–205 (2005).

13 Higbee, S., Messinger, D., Tra, Y., Voelkel, J., and Chilton, L., "A Bayesian approach to identification of gaseous effluents in passive LWIR imagery," in [*Defense, Security, and Sensing*], *Proc. SPIE*, 73341T–73341T, International Society for Optics and Photonics (2009).

14 Heasler, P., Posse, C., Hylden, J., and Anderson, K., "Nonlinear Bayesian algorithms for gas plume detection and estimation from hyper-spectral thermal image data," *Sensors* **7**(6) (2007).

15 Niu, S., Golowich, S. E., and Manolakis, D. G., "Algorithms for remote quantification of chemical plumes: a comparative study," *Proc. SPIE* **8390**, 83902I–83902I–11 (2012).

16 Goody, R. M. and Yung, Y. L., [*Atmospheric radiation: theoretical basis*], Oxford University Press, USA (1995).

17 Cha, S., "Comprehensive survey on distance/similarity measures between probability density functions," *International Journal of Mathematical Models and Methods in Applied Sciences* **1**(4), 300–307 (2007).

18 Sokolova, M. and Lapalme, G., "A systematic analysis of performance measures for classification tasks," *Information Processing & Management* **45**(4), 427–437 (2009).

19 Kraut, S., Scharf, L. L., and McWhorter, L. T., "Adaptive subspace detectors," *IEEE Transactions on Signal Processing* **49**, 1–16 (Jan. 2001).

20 Pieper, M. L., Manolakis, D., Lockwood, R., Cooley, T., Armstrong, P., and Jacobson, J., "Hyperspectral detection and discrimination using the ACE algorithm," *Proc. SPIE* **8158**, 815807–815807–12 (2011).

21 Griffin, M., Czerwinski, R., Upham, C., Wack, E., and Bruke, H., "A procedure for embedding effluent plumes into LWIR imagery," *Proc. SPIE* **5806** (2005).

22 Hammond, B. and Popa, M., "Overview of the joint services lightweight standoff chemical agent detector (JSLSCAD)," in [*Chemical and Biological Sensing VI*], *Proc. SPIE* **5795**(1), 86–95 (2005).